UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, DC 20233-0001

February 28, 2001

DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM SERIES B-13*

MEMORANDUM FOR Howard Hogan
Chief, Decennial Statistical Studies Division

From: Donna Kostanich
Assistant Division Chief, Sampling and Estimation
Decennial Statistical Studies Division

Prepared by: Alfredo Navarro and Mark Asiala
Variance Estimation and Long Form Estimation Staff

Subject: Accuracy and Coverage Evaluation: Comparing Accuracy

The attached document was prepared, per your request, to assist the Executive Steering Committee on A.C.E. Policy in assessing the data with and without statistical correction.

This report summarizes loss function analysis results comparing the accuracy of Census 2000 and the Accuracy and Coverage Evaluation.

DSSD Census 2000 Procedures and Operations Memorandum Series B-13*

February 28, 2001

Accuracy and Coverage Evaluation:
Comparing Accuracy

Alfredo Navarro and Mark Asiala

U.S. Census Bureau

Table of Contents

Tables

Accuracy and Coverage Evaluation: Comparing Accuracy

Prepared by :
Alfredo Navarro and Mark Asiala

Executive Summary

We assessed the accuracy of the census and the adjusted census (hereinafter the A.C.E.) for states, congressional districts, and sub-state areas. This involves defining and estimating error components in the A.C.E. through the Total Error Model. It also involves estimating census error or net undercount. In addition, we define a criterion by which to measure improvement from the A.C.E.

What is the Total Error Model?

The Total Error Model analyzes errors in the A.C.E. through a decomposition of the components of sampling and non-sampling errors. The error components are parameters in the model. The output from the Total Error Model is used to produce population targets by removing DSE biases. The targets are used to assess the accuracy of the A.C.E. and the census.

What are the components of error in the A.C.E.?

The components of errors in the A.C.E. are sampling variance and biases such as matching error, duplication, errors by respondents and interviewers during data collection or fictitious names in the A.C.E. Other sources of error are correlation bias and error due to ratio estimation.

What are the errors in the Census?

The census error is measured by the net undercount. The census coverage is also subject to heterogeneity across areas. This error is not quantified and is not included in the Total Error analysis.

How are the error components estimated?

Some of the estimates of component errors are based on evaluation studies of the 1990 Post-Enumeration Survey (PES) because the A.C.E. evaluation results are not available. The 1990 component error estimates were adjusted for differences in post-strata definition and changes in population size. Estimates of correlation bias were modeled based on estimates from Demographic Analysis. Sampling variances and ratio-estimator bias were calculated from the 2000 A.C.E. data.

What are loss functions?

Loss functions are scalar measures of accuracy that summarize the error in estimates. The estimates considered here are measures of population size or population shares from the census or the A.C.E. The loss functions are based on squared differences between the estimates and the targets. They differ from one another in how weighting factors are applied to summarize the results across areas.

What is the purpose of the loss function analysis?

To evaluate the relative accuracy of population counts or shares for the census and the A.C.E.

What is the measure of improvement?

The criterion we use specifies that the A.C.E. is more accurate than the census when the estimated loss for the census is greater than the estimated loss for the A.C.E. This means that an adjustment of the census based on the A.C.E. will improve the census. Loss function analysis compared the estimated loss from the census to the estimated loss from the A.C.E. for several geographic areas.

What are the units of analysis?

The loss function analysis primarily focused on population shares for congressional districts, where a unit's share was defined relative to the state population size. Concerns on numeric accuracy were addressed by examining loss function results for states and counties. A secondary concern for states is distributive accuracy because of the relevance of state shares for allocation of monies through federal programs.

What is the effect of variation in the error parameters on the loss function results?

The loss function analysis is quite sensitive to variations in the assumption of correlation bias. For congressional districts the range of improvement is from no improvement to about 1.65 (or 65 percent) when a full allowance for correlation bias is made. For example, 1.65 means that the census loss is 65 percent higher than the A.C.E. loss and therefore the A.C.E. is more accurate. Note that the results are not sensitive to the choice of correlation bias model, the census having anywhere from 1.6 to twice the amount of loss as the A.C.E. depending on the choice of correlation bias model. The results are generally favorable to the A.C.E. for all areas with respect to assumptions on the levels of processing and data collection errors.

What is the effect of correlation bias?

Under the assumption of no correlation bias the loss function results show that the census is more accurate for state numeric accuracy. For state, numeric accuracy with no correlation bias the census loss is about one-half the A.C.E. loss. The A.C.E. is more accurate for state population shares and about the same for congressional district shares. Assuming a moderate presence of correlation bias in the A.C.E. (somewhere between 20 percent and 50 percent) the loss function analysis shows the A.C.E. is more accurate for states, congressional districts, all counties, and large counties (with more than 100,000 population.)

The one exception to this finding are small counties. For numeric accuracy, under the assumption of substantial level of correlation bias present in the DSE, the weighted squared error loss function results show that the census is substantially more accurate than the A.C.E.

What is the effect of processing errors?

The analysis of the A.C.E. operations suggests that the errors were better controlled and in all likelihood are smaller in 2000 than they were in 1990. The loss function results show significant improvement from the A.C.E. as the level of processing error is reduced. The range of A.C.E. improvement is from 4.41 to 17.49 for states and 1.65 to 2.07 for congressional districts. For state levels, with 1990 levels of processing error the census loss is more than 4 times the A.C.E. loss and when processing errors are completely eliminated the census has over 17 times the A.C.E. loss. For congressional district shares, the census loss is 65 percent larger than the A.C.E. loss. For zero processing errors, the census loss is twice as much as the A.C.E. loss.

What is the effect of data collection errors?

Under assumptions of small changes in data collection error, (10 percent change in each direction) the measured improvement of accuracy of the A.C.E. relative to the census increases by 70 percent to 80 percent for both, states and congressional districts (See Tables 7 and 8).

INTRODUCTION

This memorandum summarizes and describes research undertaken by the Census Bureau to provide input into the decision whether to use the estimates of 2000 census coverage from the Accuracy of Coverage Evaluation Survey (A.C.E.) to produce the redistricting data files (P.L. 94-171.) The census counts should be considered more accurate if the corrected census leads to congressional districts within each state that are more equal in population size. The information required is the accuracy of the census, the accuracy of the corrected census numbers, and a criterion for determining which set of numbers is more accurate or has less error. The criterion proposed specifies that adjustment improves redistricting if and only if the accuracy for estimated district sizes is better for the A.C.E. than for the census. This criterion has the properties that
(i) all congressional district errors are treated approximately the same regardless of state, and (ii) a states's contribution to the overall measure of error is zero if all of the congressional districts in the state are equal in actual size, regardless of error in the state population estimate. To apply this criterion we used the existing congressional districts defined after the 1990 census. The main focus of this analysis is on congressional districts, however, state and sub-state area results are also included.

One method of evaluating the accuracy of the distribution of population estimates - shares or levels - is using loss functions. The total error simulations (Mulry and Spencer, JASA 1993) provide estimates of a target population for the analysis comparing the census and the corrected census. The Census 2000 analysis uses a combination of sources of data to estimate the distributional properties of the component errors for the 2000 A.C.E. Since the evaluations of data collection and processing errors in the 2000 A.C.E. will not be available prior to the decision, we use results from the 1990 Post-Enumeration Survey (PES) Evaluation Studies as the basis to produce estimates of these component errors. Mulry and Spencer [2] gives a detailed description of the error components and the simulation methodology.

LOSS FUNCTIONS AS MEASURE OF ERROR

Loss functions provide a conceptual framework for assessing the accuracy of population estimates (Spencer 1986, Mulry and Hogan, 1986.) Let X and T denote vectors of population estimates and their target value, respectively. The ith elements, X_i and T_i, correspond to area i. For this analysis the areas are states and congressional districts. A summary measure of the error in X as an estimate of T is obtained as the loss function L(X,T). We say that X is more accurate that an alternative estimator Y if the expected value of L(X,T) is less than the expected value of L(Y,T). The difference between L(X,T) and L(Y,T) indicates the difference between levels of accuracy.

Several criteria can be used for choosing a loss function for evaluating population estimates. For congressional districts we propose a loss function that focuses on population

shares. A unit's share is the ratio of the population of that unit to the total population of the set of units. Congressional district shares are defined relative to the state the congressional district is in. Note that our main concern is equality of congressional representation or district sizes. In this situation the unit's share is more relevant than the total.

Numeric accuracy is related to getting the count closer to the true total. Distributive accuracy pertains to getting the allocation of the population among areas closer to the true distribution. We also considered the accuracy of population shares because of the acknowledged importance of the uses of population data in allocation programs for which the share is the key. These programs are referred to as "fixed pie" allocations because the amount of benefits to be distributed is fixed a-priori.

For this analysis we use loss functions as follows:

(2.1) Congressional Districts : $L(X,T) = \sum_j Cen_j^2 \sum_i (P_{ij} - P_{Tij})^2$

Cen - Census population
P_{ij} - population share for unit i based on the census or the corrected census
P_{Tij} - target population share; i denotes congressional district i and j denotes state j

(2.2) States : $L(X,T) = \sum_j w_j (P_j - P_{Tj})^2$

P_j and P_{Tj} are defined analogously.

For most of our analysis the estimate X and its target value T are taken to be the unit's population share and this is reflected in the above definitions. Note that each of the loss functions is minimized when the population shares are perfectly estimated either by the census or the corrected census. Loss function (2.1) treats errors in all congressional districts the same regardless of the state a congressional district belongs to. The measure of accuracy should not vary with minor changes in the weight (the weight could be based on the corrected census count) and to avoid complexity the census count is used to estimate the loss for both the census and the corrected census. The weight w_j in (2.2) is conveniently taken to be inversely proportional to the estimated population share. The use of this weight has the effect of reducing the effect of large states in the overall measure of accuracy.

The loss functions provide a criterion for discriminating between the census and the corrected census on the basis of accuracy. Correcting the census with the A.C.E. results improves accuracy, and improves the quality of district sizes, if and only if the difference between the census estimated loss and the A.C.E. estimated loss is greater than zero. In the language of statistical decision theory, the measured difference is equal to the difference of the expected value of the two loss functions.

ESTIMATION

The statistical properties of the census and the corrected census estimates must be estimated. The process for doing so is referred to as "Total Error Model" or TEM. The TEM is the basis for forming an estimate of the loss for the census and the corrected census by what we call "loss function analysis". The development of the loss estimates depends on estimates of bias and variance. The methodology is briefly discussed below. For more details see Appendix 1. See Mulry and Spencer [2] for a more comprehensive discussion.

Estimation of Targets

We use data from the 1990 PES Evaluation Master Variance File (1992 PCR File) to estimate the data collection and processing errors for the 2000 A.C.E. This involved combining additional geographic information to assign each record to a 2000 A.C.E. poststratum. Calculating estimates with this file means using characteristics from the 1990 Census. This way, we have comparability by applying component errors for blocks with a high mail response rate in the 1990 Census to blocks with a high mail response rate to the 2000 Census.

Since the 1990 Evaluation Sample is not large enough to support reliable direct estimates for the 2000 A.C.E. poststrata, we first compute direct estimates for the 16 evaluation poststrata and then form model-based estimates for the poststrata. We use synthetic estimation methodology for the model-based estimation. For sampling error, imputation error, correlation bias, and ratio estimator bias, we use direct estimates for the poststrata using 2000 data.

The synthetic estimation has two phases. First, we apply the synthetic estimation to the estimates of the gross component errors to distribute the them to the seven age-sex groups within each evaluation poststratum, called the intermediate poststrata. After generating the bias from the total error simulation for the 112 intermediate poststrata, we distribute it among the poststrata within an intermediate poststratum according to the DSE and according to the absolute net undercount.

Our motivations for these choices for the synthetic estimation are: The ratios of the component errors between the age and sex groups within each minority (or nonminority) evaluation poststratum then equal the ratios for minorities (or nonminorities) at the national level. For distribution proportional to the DSE, the relative bias in the DSEs for the A.C.E. poststrata equals the relative bias in the DSE of their intermediate poststratum. For distribution proportional to the absolute net undercount, the poststratum with the largest absolute net undercount has the largest portion of the bias.

To describe the estimation for a component error, let u^+ and u^- be means of the gross errors and u be the mean of the net error such that $u = u^+ - u^-$. To calculate the synthetic estimates we first derive the direct estimates u_j^+ and σ_j^2 of a positive gross error component and

its variance in evaluation poststratum j. We also estimate the negative gross error component and its variance and follow the same derivations. In addition we estimate the covariance σ_{jkm} between the k^{th} and m^{th} error component in evaluation poststratum j. We then split the evaluation poststrata into two major groups, minority and nonminority. For each of the seven age-sex groups in each of the two major groups of evaluation poststrata, we derive direct estimates of each gross error component t_j^+. We then estimate an error component and its variance for an intermediate poststratum by $u_j^+(t_i^+/\text{sum } t_i^-)$ and $\sigma_j^2(t_i^+/\text{sum } t_i^+)$ where j denotes the corresponding evaluation poststratum and i denotes the corresponding minority or nonminority age-sex group. The covariance between the kth and mth error components in an intermediate poststratum is estimated by $\sigma_{jkm}(t_{ki}^+/\text{sum } t_{ki}^+)(t_{mi}^+/\text{sum } t_{mi}^+)$.

The simulation methodology calls for generating 1000 simulations of the coverage correction factors (CCFs). These simulations are created by generating 1000 draws from a multivariate normal distribution with mean equal to the vector of production CCFs and variance equal to the variance-covariance matrix of the production CCFs. The simulations are used to account for variance of the A.C.E. estimates and variance of the estimated A.C.E. biases.

Model bias or correlation bias is measured by comparing the A.C.E. estimates of population size with estimates of sex ratios from demographic analysis (See Bell, 1993). The difference between this estimate and the A.C.E. estimate of males is assumed to reflect model error. Demographic analysis calculates alternative estimates of the sex ratios for age and race groups. Of course, demographic analysis is subject to errors, but using the sex ratios, as opposed to the estimates of population size, remove or minimize the effect of such errors. The method used to develop estimates of correlation bias assumes no correlation bias for females. The methodology and model assumptions used to produced the estimates of correlation bias are documented in Bell, (2001.)

Estimates of Loss Functions

In the loss function analysis, the proportionate shares for the census and the corrected census numbers for each geographic area (states and congressional districts) are both compared to the proportionate shares for the target count. Using the average target to estimate the expected loss, or error, results in a biased estimate. The bias occurs because the target is an estimate and not the true value. We use a bootstrap bias correction technique to remove the bias of the estimate of expected loss for the census and the adjusted numbers. The methodology is well documented in Mulry (1992) and Navarro (1992). For a detailed presentation see the Appendix. For an additional reference see Mulry and Spencer [2].

RESULTS

We examined the effect of various assumptions about correlation bias, processing errors, and data collection errors on the loss function analysis through sensitivity analysis. This analysis was implemented by varying the assumptions underlying the estimates of component errors in the total error model. Sensitivity analysis also allows for some assessment of the robustness of the implied assumption that the Total Error Model reflects all measurable error in the A.C.E.

Correlation Bias Sensitivity Analysis

The Dual-system Estimator (DSE) contains correlation bias if any of the following assumptions are not met.

- Causal Independence - A person's participation in the A.C.E. is independent of his or her census participation.

- Homogeneity - Within post-strata, persons have the same probability of participation in the census and/or the A.C.E. Failure to this assumption leads to heterogeneity.

Model Selection

The presence of correlation bias in the A.C.E. is suggested by Demographic Analysis. For a reference see Bell (1993) and Robinson et. al., [JASA, 1993]. So, it is reasonable to assume that the DSE is subject to varying degrees of correlation bias. Estimates from Demographic Analysis (DA) are the basis to model correlation bias in the A.C.E. estimates. Unfortunately, DA estimates of correlation bias are produced only at the national level. Bell [1993, 2000] used several models to produce estimates of correlation bias. We studied the effect of the "choice of model" on the loss function analysis for states and congressional districts.

The results summarized in Table 1 show that the "choice of model" has little effect on the loss function results for congressional districts and weighted squared error on levels (numeric accuracy) for states. The numbers in the last column of Table 1 is defined as the ratio of the expected census loss to the expected A.C.E. loss. The difference between the ratio and 1 times 100 can be interpreted as the A.C.E. percent improvement. It has a somewhat more noticeable effect on weighted squared error on state shares (distributive accuracy.) All the targets are between a 5,000 (283,840,365 - 283,835,552) population range. As a result, estimates of correlation bias for all subsequent loss function analyses were based on the Two Group Model assuming no correlation bias for non-Black males 18-29 years old.

Table 1 A Effect of Choice of Correlation Bias Model on Loss Functions for States and Congressional Districts

Correlation Bias Model	Total Target Population	State Weighted Levels	State Weighted Shares	CD Relative State Share
Two Group Model except NB 18-29. Revised DA	283,837,998	4 895	1.793	1 648
Fixed Odds Ratio Model except NB 18-29, Revised DA	283,840,365	5.121	2 179	1.815
Fixed Ratio of PM22 to PF22 Model except NB 18-29, Revised DA	283,841.734	5 033	1.987	1.746
Fixed Relative Risk Model except NB 18-29. Revised DA	283,838.451	5 023	2 009	1 758
Generalized Behavior Response Model except NB 18-29, Revised DA	283,835.552	4 897	1.950	1.797
Prithwis Das Gupta's Model except NB 18-29. Revised DA	283,838,808	4 785	1 592	1 561

The weighted squared error loss function for levels is used to ascertains numeric accuracy. It measures the error in the population counts as opposed to the population shares.

How much correlation bias is needed to show an improvement from the A.C.E.?

Under the assumption of no correlation bias in the DSE and other errors such as processing and data collection at the 1990 levels, the census and the A.C.E. are equally accurate for congressional districts (Census Loss / A.C.E. Loss = .995, Table 3.B). For states, the results are mixed without correlation bias. The ratios for the weighted squared error on levels and shares are 0.519 and 1.783, respectively. However, assuming correlation bias is zero is clearly wrong. There is strong evidence from DA that correlation bias is present in the DSE, especially for adult Black males.

Tables 2,3, and 4 summarize the results for congressional districts, counties and states, respectively. We considered 4 rates of correlation bias: 10, 20, 50, and 75 percent.

- Under the assumption of full correlation bias and error levels (matching and data collection errors) similar to 1990, the evidence is very strong in favor of improvement from the A.C.E., that is, the A.C.E. is decisively more accurate than the census for population shares. The percent improvements are 65 and 78 for congressional districts and states, respectively (see Table 3.B).

- Assuming very modest reduction in the level of processing error (10 percent reduction), loss function results are very sensitive to the assumption of correlation bias. For population shares, the results indicate that the A.C.E. is more accurate for all units of

analysis even for small levels (10 - 20 percent) of correlation bias. For levels the results are mixed. The census is more accurate for states and counties for small levels of correlation bias. As the degree of correlation bias in the DSE increases - 50 percent and higher - the A.C.E. becomes more accurate than the census for states and counties. The census is more accurate than the A.C.E. for counties with less than 100,000 population. For this analysis (Table 3) the focus is numeric accuracy and counties are grouped within size categories. This finding suggest that for these areas any improvement from adjustment is offset by the A.C.E. variance and additional biases. The A.C.E. and the census are equally accurate for large counties (100,000 population or more) assuming little or no correlation bias. Assuming less than strong presence of correlation bias in the DSE the A.C.E. is more accurate than the census for large counties.

- Assuming similar correlation bias for Hispanics and Blacks, and assuming zero correlation bias for Owners suggest that the A.C.E. is more accurate than the census. (See Table 1.B, below.)

Table 1.B Additional Analysis for Owners and Hispanics

Correlation Bias Model	Total Target Population	Weighted Levels (States)	Weighted Shares (States)	CD Relative State Share
Correlation Bias except NB 18-29 and All Owners, Revised DA	283,139,516	1.276	1.908	1.578
Correlation Bias except NB 18-29, Hispanic same as Black Corrected, Revised DA	284,191,614	10.922	2 326	2 082

Table 2 Effect of Correlation Bias on Loss Functions for Congressional Districts

Correlation Bias Model	Total Census Population	Total Actual ACE Population	Total Sim ACE Population	Total Target Population	CD Relative State Share
10% Correlation Bias Black Only, 90% Processing Error, Revised DA	281,421,906	284,683,787	284,678,060	282,901,001	1.147
20% Correlation Bias Black Only, 90% Processing Error, Revised DA	281,421,906	284,683,787	284,678,060	282,975,111	1.265
50% Correlation Bias Black Only, 90% Processing Error, Revised DA	281,421,906	284,683,787	284,678,060	283,198,142	1 554
75% Correlation Bias Black Only, 90% Processing Error, Revised DA	281,421,906	284,683,787	284,678,060	283,383,776	1 686

Total Sim ACE Population is the average total population across the 1000 simulations of the A.C.E. Total Target Population is the average total population across the 1000 simulations of the target population.

Table 3.A. Effect of Correlation Bias on Loss Functions for Counties

Geography	Correlation Bias Model	Total Census Population	Total Actual ACE Population	Total Sim ACE Population	Total Target Population	Weighted Levels	County Relative State Share
County	10% Correlation Bias Black Only, 90% Processing Error, Revised DA	281,421,906	284,683,795	284,678,082	282,900,984	0 748	1.228
County	20% Correlation Bias Black Only, 90% Processing Error, Revised DA	281,421,906	284,683,795	284,678,082	282,975,107	0.853	1.361
County	50% Correlation Bias Black Only, 90% Processing Error, Revised DA	281,421,906	284,683,795	284,678,082	283,198,112	1.243	1.703
County	75% Correlation Bias Black Only, 90% Processing Error, Revised DA	281,421,906	284,683,795	284,678,082	283,383,743	1.643	1 881
County <= 100,000 Pop	10% Correlation Bias Black Only, Revised DA	69,489,081	70,186,846	70,185,741	69,506,988	0.155	N/A
County <= 100,000 Pop	20% Correlation Bias Black Only, Revised DA	69,489,081	70,186,846	70,185,741	69,519,363	0.162	N/A
County <= 100,000 Pop	50% Correlation Bias Black Only, Revised DA	69,489,081	70,186,846	70,185,741	69,556,536	0 196	N/A
County <= 100,000 Pop	75% Correlation Bias Black Only, Revised DA	69,489,081	70,186,846	70,185,741	69,587,514	0 240	N/A
County > 100,000 Pop	10% Correlation Bias Black Only, Revised DA	211,932,825	214,496,949	214,492,341	213,264,482	1.005	N/A
County > 100,000 Pop	20% Correlation Bias Black Only, Revised DA	211,932,825	214,496,949	214,492,341	213,326,249	1.176	N/A
County > 100,000 Pop	50% Correlation Bias Black Only, Revised DA	211,932,825	214,496,949	214,492,341	213,512,062	1 836	N/A
County > 100,000 Pop	75% Correlation Bias Black Only, Revised DA	211,932,825	214,496,949	214,492,341	213,666,708	2.534	N/A

Table 3 B Relative Loss by Degree of Processing Error and Correlation Bias

Model	Degree of Correlation Bias	Degree of Processing Error	Census Loss / A.C.E. Loss (St. Levels)	Census Loss / A.C.E. Loss (St. Shares)	Census Loss / A.C.E. Loss (CD shares)
NA	0%	100%	0 519	1.783	0 995
1	100%	0%	17 488	1 125	2.068
1	100%	25%	18.565	1.318	1 975
1	100%	50%	14 108	1 500	1.870
1	100%	75%	8 242	1 656	1.759
1	100%	100%	4 413	1.780	1.651
2	10%	90%	0.770	1 761	1.147
2	20%	90%	0.897	1.792	1.265
2	50%	90%	1.416	1.838	1.554
2	75%	90%	2.048	1.821	1.688

Model 1 - correlation bias is present for males except for Non-black males age 18 to 29.
Model 2 - correlation bias is present for Black males only.
States use weighted squared error loss and congressional districts use equal CD squared error loss.

Table 4 Effect of Correlation Bias on Loss Functions for States

Correlation Bias Model	Total Census Population	Total Actual ACE Population	Total Sim ACE Population	Total Target Population	Weighted Levels	Weighted Shares
10% Correlation Bias Black Only, 90% Processing Error, Revised DA	281,421,906	284,683,794	284,678,078	282,900,999	0 770	1.761
20% Correlation Bias Black Only, 90% Processing Error, Revised DA	281,421,906	284,683,794	284,678,078	282,975,113	0 897	1.792
50% Correlation Bias Black Only, 90% Processing Error, Revised DA	281,421,906	284,683,794	284,678,078	283,198,131	1.416	1 838
75% Correlation Bias Black Only, 90% Processing Error, Revised DA	281,421,906	284,683,794	284,678,078	283,383,781	2.048	1 821

Processing Error Sensitivity Analysis

The A.C.E. made a significant number of design improvements in order to reduce biases associated with the 1990 Post-Enumeration Survey. The A.C.E. design included a much improved matching process. Automation was built into the system in all phases. Edits and quality checks were built into the system to reduce the 1990 levels of error. In addition the matching process was completely centralized in one site as opposed to be decentralized as in 1990. This facilitates the implementation of effective controls and a more uniform application of the matching rules. Therefore it is reasonable to expect at least modest gains in the level of processing error (mostly matching error) compared to 1990. We varied the reduction of processing errors from no improvement to 100 percent (assuming correlation bias except for non-Black males 18-29 years of age) and implemented a sensitivity analysis for congressional districts and states. The results are summarized in Tables 5 and 6.

- As expected, as the processing error becomes smaller the measured improvement of accuracy of the A.C.E. relative to the census increases for congressional districts and states for levels.

- For states, the weighted squared error on shares shows an opposite trend, that is, as the level of processing error reduces the measured improvement of accuracy of the A.C.E. relative to the census decreases but the A.C.E. still appears more accurate than the census.

Table 5 Effect of Processing Error on Loss Functions for Congressional Districts

Correlation Bias Model	Total Census Population	Total Actual ACE Population	Total Sim ACE Population	Total Target Population	CD Relative State Share
Correlation Bias except NB 18-29, 0% Processing Error	281,421,906	284,683,787	284,678,060	285,088,509	2.068
Correlation Bias except NB 18-29, 25% Processing Error	281,421,906	284,683,787	284,678,060	284,761,146	1.975
Correlation Bias except NB 18-29, 50% Processing Error	281,421,906	284,683,787	284,678,060	284,434,845	1.870
Correlation Bias except NB 18-29, 75% Processing Error	281,421,906	284,683,787	284,678,060	284,110,248	1.759
Correlation Bias except NB 18-29, 100% Processing Error	281,421,906	284,683,787	284,678,060	283,785,901	1 651

Table 6 Effect of Processing Error on Loss Functions for States

Correlation Bias Model	Total Census Population	Total Actual ACE Population	Total Sim ACE Population	Total Target Population	Weighted Levels	Weighted Shares
Correlation Bias except NB 18-29, 0% Processing Error	281,421,906	284,683,794	284,678,078	285,088,512	17.488	1.125
Correlation Bias except NB 18-29, 25% Processing Error	281,421,906	284,683,794	284,678,078	284,761,136	18.565	1.318
Correlation Bias except NB 18-29, 50% Processing Error	281,421,906	284,683,794	284,678,078	284,434,836	14.108	1 500
Correlation Bias except NB 18-29, 75% Processing Error	281,421,906	284,683,794	284,678,078	284,110.255	8 242	1 656
Correlation Bias except NB 18-29, 100% Processing Error	281,421,906	284,683,794	284,678,078	283,785,900	4 413	1.780

Data Collection Errors Sensitivity Analysis

The A.C.E. used computer laptops to conduct Computer Assisted Person Interviewing rather than the paper instrument used in 1990. We also implemented several improvements into system design and software development to reduce the risk of computer processing errors and increase data quality. Therefore, it is reasonable to assume at least modest gains in data collection accuracy compared to 1990. We studied the effect of this assumption on the loss function results for congressional districts and states by simulating the loss function for a 10 percent two-way change compared to the 1990 PES. In addition, a 10 percent reduction in processing error is assumed. We studied these two effect for several levels of correlation bias. Results are summarized in Tables 7 and 8.

- For congressional districts, the A.C.E. is more accurate even when the data collection accuracy is increased by 10 percent.

- Reducing the level of data collection error by little improves states numeric accuracy with the A.C.E. even under little presence of correlation bias in the DSE (from .744 to .884 for 10 percent correlation bias and .853 to 1.029 for 20 percent correlation bias - See Tables 4 and 8) , for larger levels of correlation bias the results are obviously favorable to the A.C.E.

Table 7 Effect of Data Collection Error on Loss Functions for Congressional Districts

Correlation Bias Model	Total Census Population	Total Actual ACE Population	Total Sim ACE Population	Total Target Population	CD Relative State Share
10% Correlation Bias Black Only, 90% Processing Error, 90% Data Collection Error, Revised DA	281,421,906	284,683,787	284,678,060	282,963,842	1.169
20% Correlation Bias Black Only, 90% Processing Error, 90% Data Collection Error, Revised DA	281,421,906	284,683,787	284,678,060	283,038,025	1.289
50% Correlation Bias Black Only, 90% Processing Error, 90% Data Collection Error, Revised DA	281,421,906	284,683,787	284,678,060	283,261,305	1.580
75% Correlation Bias Black Only, 90% Processing Error, 90% Data Collection Error, Revised DA	281,421,906	284,683,787	284,678,060	283,447,267	1 708
10% Correlation Bias Black Only, 90% Processing Error, 110% Data Collection Error, Revised DA	281,421,906	284,683,787	284.678,060	282,838,205	1.125
20% Correlation Bias Black Only, 90% Processing Error, 110% Data Collection Error, Revised DA	281,421,906	284,683,787	284,678,060	282,912,237	1.240
50% Correlation Bias Black Only, 90% Processing Error, 110% Data Collection Error, Revised DA	281,421,906	284,683,787	284,678,060	283,135,064	1 528
75% Correlation Bias Black Only, 90% Processing Error, 110% Data Collection Error, Revised DA	281,421,906	284.683,787	284,678,060	283,320,598	1.664

Table 8 Effect of Data Collection Error on Loss Functions for States

Correlation Bias Model	Total Census Population	Total Actual ACE Population	Total Sim ACE Population	Total Target Population	Weighted Levels	Weighted Shares
10% Correlation Bias Black Only, 90% Processing Error, 90% Data Collection Error, Revised DA	281,421,906	284,683,794	284,678,078	282,963,840	0.884	1.844
20% Correlation Bias Black Only, 90% Processing Error, 90% Data Collection Error, Revised DA	281,421,906	284,683,794	284,678,078	283,038,029	1.029	1.874
50% Correlation Bias Black Only, 90% Processing Error, 90% Data Collection Error, Revised DA	281,421,906	284,683,794	284,678,078	283,261,300	1 619	1.910
75% Correlation Bias Black Only, 90% Processing Error, 90% Data Collection Error, Revised DA	281,421,906	284,683,794	284,678,078	283,447,269	2 333	1 882
10% Correlation Bias Black Only, 90% Processing Error, 110% Data Collection Error, Revised DA	281,421,906	284,683,794	284,678,078	282,838,206	0 669	1 677
20% Correlation Bias Black Only, 90% Processing Error, 110% Data Collection Error, Revised DA	281,421,906	284,683,794	284,678,078	282,912,239	0 780	1.709
50% Correlation Bias Black Only, 90% Processing Error, 110% Data Collection Error, Revised DA	281,421,906	284,683,794	284,678,078	283,135,062	1 237	1.763
75% Correlation Bias Black Only, 90% Processing Error, 110% Data Collection Error, Revised DA	281,421.906	284,683,794	284,678,078	283,320,603	1.796	1 757

References

Mulry, M. H., and Spencer, B. D. (1993), "Accuracy of the 1990 Census and Undercount Adjustments", Journal of the American Statistical Association, 88, 1080-1091.

Mulry, M. H., and Spencer, B.D., (2001), "Overview of Total Error Modeling and Loss Function Analysis," DSSD Census 2000 Procedures and Operations Memorandum Series B-19*, February 28, 2001.

Spencer, B. D. , (1986), "Conceptual Issues in Measuring Improvement in Population Estimates", ARC Proceedings, 393-407.

Mulry, M. H., and Hogan, H., (1986), "Research Plan on Census Adjustment Standards", ARC Proceedings, 381 - 392.

Mulry, M. H., (1992), "Loss Function Analysis for the Post Census Review (PCR) Estimates", Unpublished Census Bureau Memorandum.

Navarro, A., (1992), "Additional Loss Function Analysis - Technical Report", Unpublished Census Bureau Memorandum.

Bell, W. R., (1993), "Using Information from Demographic Analysis in Post_Enumeration Survey Estimation" 88, 1107-1118.

Robinson, J. G., (1993), "Estimation of Population Coverage in the 1990 United States Census Based on Demographic Analysis", 88, 1061-1079.

Appendix - Loss Function Calculation

This appendix summarizes the way calculations of loss functions were performed and explains the logic behind the calculations. The logic of the analysis is fairly straightforward, but is easily lost in the trees. To explain it we use some simple notation, which will be replaced by more complex notation when the details are described, below. (Also see section 1, above.) Let C denote the census estimate, A the adjusted estimate, and B an estimate of bias in the adjusted estimate. Let V_A denote an estimate of variance of A and let V_B denote an estimate of variance of B; we assume A and B have negligible correlation. To estimate the mean squared error (MSE) of C and A we construct a "target" estimate, T, defined as T = A - B. If T had zero variance, we could estimate the MSEs by $(C - T)^2$ and $(A - T)^2$. The variance of T is approximately $V_A + V_B$, however, and so we estimate the MSE of C by

$$(C - T)^2 - (V_A - V_B) \tag{1}$$

and we estimate the MSE of A by

$$B^2 - V_A - V_B. \tag{2}$$

The excess MSE of C relative to the MSE of A is estimated by $(C - T)^2 - B^2 - 2V_A$; observe that the specification for V_B does not affect point estimates of the difference in the MSEs.

The variances are calculated by means of replicates. The basis for the calculation of adjusted estimates and targets consists of (i) the vector of adjustment factors for poststrata, (ii) the estimated covariance matrix of the adjustment factors, (iii) the vector of estimated biases of the adjustment factors, and (iv) the estimated covariance matrix of the estimated biases. The vectors of replicates are constructed by random sampling from a multivariate normal distribution with covariance matrix equal to the estimated covariance matrix of (i) or (iii), as the case may be. To estimate the variance of a function of (i) or (iii), we calculate the function for each replicate and use the empirical variance among the calculated values.

Notation

Subscript h ($1 \le h \le H$) will refer to poststrata and the subscript i ($1 \le i \le I$) will refer to general areas such as states, counties, cities, congressional districts, etc. The subscripts q ($1 \le q \le Q$) and s ($1 \le s \le S$) will be used to denote replicates. The replicates are constructed so that their empirical variance over q provides an estimate of variance due to random sampling (and, perhaps, imputation) in the DSE and their empirical variance over s provides an estimate of variance due to random sampling in the evaluation studies for estimating bias in the DSE; details

are provided below. A "+" in place of a subscript denotes a total obtained by summation over that subscript. The subscript a denotes an empirical estimate and the subscript t denotes a target.

The notation is consistent with that of some other Census Bureau documentation of the calculations, except that F is used in place of AF to indicate adjustment factor; some additional notation is introduced as well. In operation, Q = S = 1000.

Census Estimates

N_{ci} census count, area i

N_{chi} census count, part of poststratum h that is in area i

N_{c+} census count for aggregation of areas

$$N_{c+} = \sum_{i} N_{ci}$$

P_{ci} population share of area i; $P_{ci} = N_{ci}/N_{c+}$

Adjusted Estimates

F_{aqh} replication q of adjustment factor for poststratum h

$$\bar{F}_{a+h} = \sum_{q=1}^{Q} F_{aqh}/Q$$

F_{ah} production adjustment factor for poststratum h

$$F_{ah} = \bar{F}_{a+h}$$

$\hat{V}_{ahi}$ estimated covariance between F_{ah} and F_{ai}

$$\hat{V}_{ahi} = \sum_{q=1}^{Q} (F_{aqh} - \bar{F}_{a+h})(F_{aqi} - \bar{F}_{a+i})/(Q-1)$$

X_{aiq} replication q of adjusted count for area i

$$X_{aiq} = \sum_{h} F_{aqh} N_{cih}$$

X_{a+q} replication q of adjusted count for an aggregation of areas

$$X_{a+q} = \sum_{i} X_{aiq}$$

X_{ai} production adjusted count for area i.

$$X_{ai} = \sum_{h} F_{ah} N_{cih}$$

X_{a+} adjusted count for an aggregation of areas

$$X_{a+} = \sum_{i} X_{ai}$$

P_{aiq} replication q of adjusted population share of area i; $P_{aiq} = X_{aiq}/X_{a+q}$

$\bar{P}_{ai+}$ average of replicates of adjusted share of area i; $\bar{P}_{ai} = \sum_{q=1}^{Q} P_{aiq}/Q$

P_{ai} production adjusted population share of area i; $P_{ai} = X_{ai}/X_{a+}$

$\hat{V}_{Pai}$ estimate of variance of P_{ai}

$$\hat{V}_{Pai} = \sum_{q=1}^{Q} (P_{aiq} - \bar{P}_{ai.})^2/(Q-1)$$

Targets

F_{tsh} replication s of target adjustment factor for poststratum h

$$\bar{F}_{t.h} = \sum_{s=1}^{S} F_{tsh}/S$$

F_{th} target estimate of adjustment factor for poststratum h

$$F_{th} = \bar{F}_{t.h}$$

$\hat{V}_{thu}$ estimated covariance between F_{th} and F_{tu}

$$\hat{V}_{thu} = \sum_{s}^{S} (F_{tsh} - \bar{F}_{t.h})(F_{tsu} - \bar{F}_{t.u})/(S-1).$$

X_{tis} replication s of target count for area i

$$X_{tis} = \sum_{h} F_{tsh} N_{cih}$$

X_{t+s} replication s of target count for an aggregation of areas

$$X_{t.s} = \sum_{i} X_{tis}$$

P_{tis} replication s of target population share of area i; $P_{tis} = X_{tis}/X_{t+s}$

$\bar{P}_{ti}$ average of replicates of target share of area i; $\bar{P}_{ti.} = \sum_{s=1}^{S} P_{tis}/S$

P_{ti} target share of area i; $P_{ti} = \bar{P}_{ti.}$

B_{Pi} estimate of bias in adjusted share, P_{ai}

$$B_{Pi} = P_{ai} - P_{ti}$$

$\hat{V}_{BPi}$ estimate of variance of B_{Pi}

$$\hat{V}_{BPi} = \sum_{s=1}^{S} (P_{tis} - \bar{P}_{ti})^2/(S-1)$$

Loss Function Calculations

First consider the MSE for the census. Define

$$L_{ci1} = (P_{ci} - P_{ti})^2$$

$$L_{cisq1} = [P_{ci} - P_{tis} + (P_{ai} - P_{aiq})]^2$$

$$\bar{L}_{ci\cdot\cdot 1} = \sum_{s=1}^{S} \sum_{q=1}^{Q} L_{cisq1}/(SQ)$$

and observe that

$$\bar{L}_{ci\cdot\cdot 1} = L_{ci1} + (1 - S^{-1})\hat{V}_{BPi} + (1 - Q^{-1})\hat{V}_{Pai}.$$

Thus,

$$2L_{ci} - \bar{L}_{ci\cdot\cdot 1} = L_{ci1} - (1 - S^{-1})\hat{V}_{BPi} - (1 - Q^{-1})\hat{V}_{Pai},$$

as desired in (1) except for the small terms in S^{-1} and Q^{-1} (see overview), and so we estimate the MSE in the P_{ci} by $L^{R}_{ci1} = 2L_{ci} - \bar{L}_{ci\cdot\cdot 1}$.

Turning attention to the adjusted estimates, define

$$L_{aiq1} = (P_{aiq} - P_{ti})^2$$

$$L_{aisq1} = (P_{aiq} - P_{tis})^2$$

$$\bar{L}_{ai\ 1} = \sum_{q=1}^{Q} L_{aiq1}/Q$$

$$\bar{L}_{ai\ \ 1} = \sum_{s=1}^{S} \sum_{q=1}^{Q} L_{aisq1}/(SQ)$$

and observe that

$$\bar{L}_{ai\cdot 1} = B^2_{Pi} - (1 - Q^{-1})\hat{V}_{Pai}$$

and

$$\bar{L}_{a1\,.1} = \bar{L}_{a1.1} - (1 - S^{-1})\hat{V}_{BPi}.$$

Thus, $2\bar{L}_{a1.1} - \bar{L}_{a1\,1} = B^2_{Pi} - \hat{V}_{Pai} - \hat{V}_{BPi}$, as desired in (2), and so we estimate the MSE of P_{a1} by $L^R_{a11} = 2\bar{L}_{a1.1} - \bar{L}_{a1\,1}$.

Notes

Error from choice of imputation method was not reflected in $\hat{V}_{ah.}$. It was reflected in $\hat{V}_{BPi}$, but that does not affect the point estimates of difference in expected loss. The variance of the estimate of correlation bias is not reflected in $\hat{V}_{BPi}$.